Clinical Documents: Attribute-Values Entity Representation, Context, Page Layout And Communication

Christian Lovis MD MPH, Alexander Lamb, Robert Baud PhD, Anne-Marie Rassinoux PhD, Paul Fabry, Antoine Geissbühler MD

Division of Medical Informatics, University Hospitals of Geneva, Geneva, Switzerland

ABSTRACT

This paper presents how acquisition, storage and communication of clinical documents are implemented at the University Hospitals of Geneva. Careful attention has been given to user-interfaces, in order to support complex layouts, spell checking, templates management with automatic prefilling in order to facilitate acquisition. A dual architecture has been developed for storage using an attributes-values entity unified database and a consolidated, patient-centered, layout-respectful files-based storage, providing both representation power and sinsert(peed of accesses. This architecture allows great flexibility to store a continuum of data types from simple type values up to complex clinical reports. Finally, communication is entirely based on HTTP-XML internally and a HL-7 CDA interface V2 is currently studied for external communication. Some of the problem encountered, mostly concerning the typology of documents and the ontology of clinical attributes are evoked.

INTRODUCTION

Managing clinical documentation efficiently for the computerized patient record remains a challenge that is even emphasized in a multi-tier, component based architecture where communication and integration are essential. In addition to the characteristics of a clinical document as proposed by Dolin [1], we had several requirements to meet:

- efficient data acquisition, whether it was a word processing program, such as used by most of our secretaries or whether it was very structured questionnaires;
- support for complex layouts used in the various clinical documents, with headers, footers, columns, tables, symbols, etc…;
- versatile storage, including the possibility to run complex queries, have alerts, but also very fast accesses for the use by clinicians in their daily work and support for heterogeneous file formats;
- data should be as structured as feasible with the finest granularity available but the possibility to have full text and narratives must be kept;
- open for communication for a future community-based or national patient record,
- support for legal and ethical regulations,
- audit trail and workflow management.

Since 1994, we use a template-based clinical document management system part of the Diogene system that allows documents to be edited and stored in a paragraph-oriented structure. Our experiment is based on the Diogenes 2 architecture [2] that allows us to have a centralized repository for various kinds of clinical narratives, including complex discharge letters. More than 900 users in our Hospital use this system in 40 medical services. All inpatient clinics are using the system for a coverage exceeding 80% of official reports. This includes radiology reports, pathology, surgical procedures and discharge letters amongst others. Progress notes, though supported by the system, are only used on a limited scale, mostly because residents have to write themselves. Most medical outpatient clinics are in the process of using the system for patient summaries and discharge reports. At the time of writing this paper, over 2 millions structured documents were available online. Among them, there are approximately 1’260’000 reports and discharge letters. Almost 7’000’000 structured paragraphs have been generated during this process. Documents are stored in two formats. On one side, they are stored as sets of paragraphs linked in a relational SQL database. This means that a “rebuild” process must be achieved in order to restore the complete document. During this process, the page layout is lost and only content or multimedia links are preserved. Therefore, all documents are also stored as read-only viewable documents in a document management system that includes a versioning management. In this system, all documents of a given patient are stored in one folder in their original format. Retrieval and display is very fast with full preservation of the documents layouts. This double storage system is the cornerstone of the clinical document and clinical attributes management within the EPR and is based on an n-tiers XML-compliant architecture. The most important evolution that will be done this year will be the implementation of interfaces using the new release 2 of the HLA Clinical

Document Architecture (CDA) made available in January 2003 [3] for external communication of these documents and the clinical attributes or paragraphs they contain in a structured manner. The HL7-CDA is a document markup standard that specifies the structure and semantics of *clinical documents* for the purpose of exchange. According to HL7-CDA, a clinical document contains observations and services and has the following characteristics [1]:

- *Persistence* – A clinical document continues to exist in an unaltered state, for a time period defined by local and regulatory requirements.
- *Stewardship* – A clinical document is maintained by a person or organization entrusted with its care.
- *Potential for authentication* - A clinical document is an assemblage of information that is intended to be legally authenticated.
- *Wholeness* - Authentication of a clinical document applies to the whole and does not apply to portions of the document without the full context of the document.
- *Human readability* – A clinical document is human readable.

All these characteristics are fundamental attributes of a *document* in order to preserve its significance within a clinical context as opposed to disparate or discrete clinical observations. A CDA document is a well-defined and complete information object that can include text, images, sounds, and other multimedia content. It is made of an XML file and eventually style sheets for simple layouts. If the CDA is very suitable for exchanging documents, edition of CDA documents is not easy when using complex page layouts and storage in its original XML format is not efficient for queries.

In real practice, most documents used in clinical settings are made both of typed or structured data and narratives or free texts. Some of them can be structured with a set of paragraphs, such as a discharge letter, others have very typed fields, such as a prescription. Most of the documents typed in our hospitals are dictated and typed by typists or secretaries. They make a strong use of word processors, complex page layouts and spell checking. It is therefore of prime importance to be able to converge in any situation towards structured documents that can be handled in such conditions.

We do present a concept that permits the unification of semi-structured information such as what can be found in free texts and strictly typed data such as fields in questionnaires, allowing a common representation at several levels, from user interfaces to data storage as it is currently being used in our hospitals.

The ultimate goal of documentation is to provide accurate and timely clinical information for patient care and complete documentation for all stakeholders [4]. The need for resource management and integrated clinical pathways requires a transversal understanding both of care structures and of data representation, management and acquisition. The main challenge is the ability to represent the knowledge in a way that is usable, maintainable and meaningful to diverse users and automatic processes. However, this implies numerous requirements that are sometimes in contradiction. To this respect, we face two major problems for the handling of structured clinical documents:

- The hierarchy (or classification) of document types (CDA `document_type_cd`). Despite the fact that there has been work on document ontologies, mostly by the HL7 Document Ontology Task Force[1] [5], no final release is available. Working with a large number of document types is almost not possible for clinicians without a good ontology, mostly in order to be able to give them pertinent views of the documents. In order to estimate the document types used in our hospital, we manually analyzed 10 paper-based patient records in each of the 32 main medical Departments, representing a total of 15'629 pages. We identified 686 different document types, Only 37 documents do appear in more than 90% of the record, and 250 document types do appear only in one record. Based on that, we establish a first release of a document ontology that we hope to replace with an international one as soon as available.

- The representation of the clinical elements being described (such as *observations* or *paragraphs*) and stored in the attribute-value entity database. The CDA body lacks a robust semantic for the full representation of a clinical fact's context. The underlying problem is far from trivial and major consequences may have to be faced in the long term. At the present time, we have no such good representation.

This paper focuses on the data-model and the storage architecture. On one side, the data-centric database that is based on an attribute-value entity architecture and defines the complete documents' content. On the other side, we use a patient-centric and file oriented

[1] Document Ontology Task Force, hl7.org/special/dotf/dotf.htm

database that contains the same documents, but with their original layouts.

MERGING QUESTIONNAIRES AND DOCUMENTS

In a move to merge document-based medical narratives and questionnaires data acquisition, we consider that questionnaires and free texts documents have common structure and representation; they are built upon a description of fields. Questionnaires use mostly basic attributes types, such as Boolean, dates,

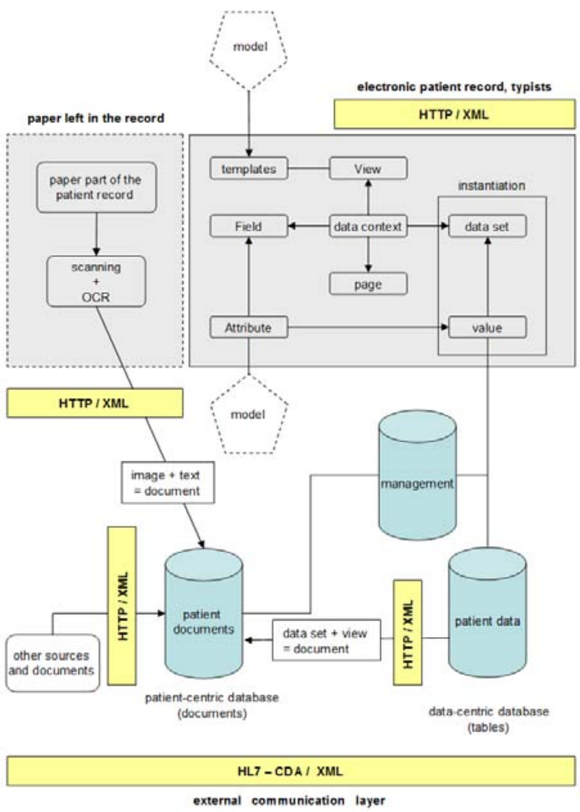

Figure 1: Overall architecture

lists or numeric whereas documents are mostly formed using paragraphs. Documents have therefore been structured based on their paragraphs, such as patient history, discussion, conclusions. These paragraphs can be made of basic attributes such as found in questionnaires or they can be made of small portions of free text [6]. All documents and questionnaires do share a set of similar characteristics used for document management such as workflow, privacy control and versioning beside the fact that they share the same attributes and templates representation.

DATA MODEL

The data model is based on the fact that all documents are built using a set of attribute-value entities (*the data fields*) grouped within a document class (*the data context*). All documents are built using a set of shared attributes with a given type, meaning and possible values. Therefore, the fields of a document are only references to attributes. The values are represented as lists of attribute-value pairs connected to an instance of a document for a given patient at a given date. Once filled, the resulting values will be stored in a separate repository, which contains attributes-values pairs.

- *Data context (or document)*
  Logical model that regroups fields. It is, in fact, the list of all fields needed to create a questionnaire, a document or a specialized record such as the record of anesthesiology consultation.
- *Field (or attribute caption)*
  Label of an attribute in a questionnaire or a document. Each field is linked to a specific attribute. For example, in a given questionnaire, it might be a field "Smoker?" linked to the attribute tobacco_use
- *Attribute (or clinical fact)*
  An attribute is one entry in the list of what can be expressed. For each attribute classes, several properties can be specified, such as data type, description, links to external models, etc. Within the dictionary, each attribute has a unique internal identifier that cannot be removed. Each fact can have dates of validity, so that facts are never deleted but only inactivated. Each attribute belongs to one of the seven basic data types, which are enumerated, date, decimal, external link, integer, long text and short text. The values that can take an attribute can optionally be limited by a code value, such as "Yes", "No". If needed, these possible values can be aggregated in Groups. Attributes can be linked to external classification, and numerous are already linked, typically to ICD10 or ICPC.
- *View (or document template)*
  A view is a way a data context will be displayed. It does not define which fields (and linked attributes) are used, but the format, layout, authorization schemes and components used for display. An example of view is PDF for Adobe Acrobat Portable Document Format®. Another of the Views of a data context could be a Microsoft Rich Text

Format template that can allow complex editing and layout.

- *Page*
  Page allows to group fields and attributes in a clinically pertinent manner. It can be used to produce automatically acquisition user interfaces. A typical Page might contain attributes pertaining to cardiology examination.

Once a *data context* has been instantiated, such as a questionnaire or a document for example, *attributes* will receive *data* (their values) either automatically or entered by users. All *data* of a *data context* represent its *data set*. A link is maintained between *data* and *attribute* to allow *attribute-values* to be retrieved, as well as between *data set* and *data context*, so that the whole context in which data have been acquired is kept.

SEMANTIC MODEL FOR ATTRIBUTES

As already evoked in the introduction, there is a need for a model of attributes, and this is a true challenge. Such a semantic model, acting as a "semantic" shield over the list of attributes, has rapidly proven to be necessary with the increase of the size of the number of attributes and the apparition of synonyms or duplicates. One of the main problems encountered has been to have a way to create rapidly new attributes without loosing the added-value of a semantic representation and without spending too much time to decide where a new attribute must be put in a model. There is a large pressure coming from users for having new attributes fast, whereas organizing these attributes in a semantic model can take long time and discussions. The dictionary of attributes can be considered as a flat list, or almost equivalent and we are building a completely separated model, used to organize and consolidate the "meaning" of attributes. The process of building this model is still ongoing. In order to avoid rebuilding a deep and complex model, we decided to have a *light model* that is manageable by human [7]. Each attribute can be linked to *0..n* concepts in the light model with only four type of links:

- *isA*. It is a subsumption relation that can be used in various cases, such as *femur isA bone* or *headache isA pain*.
- *partOf*. It is a partitioning relation that can express, for example, that *finger* belongs to the *hand*.
- *equiv*. It is an equivalence relation reserved to express synonymy or medical equivalence.
- *isNot* is a negation that can be used to ease the matching of similar concept, but that would have been expressed using negation in the attributes list.

The same links can be used to link concepts between them. Despite the fact that such a model is probably insufficient for deep analysis or real knowledge representation, it seems to be enough for the purpose of clinical documentation such as pre-filling documents with existing values, linking similar attributes between documents, etc. It offers also the advantage of having only those concepts that are used within structured documents.

STORAGE

Despite its significant benefits, entity-attributes-values design has the disadvantage to be less efficient than conventional database when accessing data. In particular, attribute-centered queries, where the query criterion is based on the value of a particular attribute, are most likely to show impaired performance [8]. The loss in performance can be even more pronounced when large document based on multiple EAV have to be rebuilt on the fly for displaying. To overpass this problem, all documents with their original layout are stored in a file-oriented database, with extremely fast access performances. The storage is made in the two main repositories (in darker in figure 1) of the CPR, the patient-centric database and the data-centric database. All accesses to the data are made through the middleware (in lighter in figure 3) using HTTP / XML messages. The two databases share a common set of components for management, accesses, auditing, etc. All information to produce valid CDA documents is present and we will be able to communicate documents using HL7 CDA as soon as a final version will be released.

Beside questionnaires and documents, the data-centric database does hold all structured data on clinical activity about patients, such as order entry and laboratory, and they can therefore be used to automatically fill documents. The patient-centric (document) database is much more heterogeneous and can store any kind of file, such as, for example, publications related to a given patient found in the medical literature and considered important by the care team. In addition, the scanning system that will be used in the near future to scan all documents still on paper will send all its outputs, scanning and texts from optical character recognition (OCR) to the patient-centric database providing a complete paperless patient record made of heterogeneous data and sources.

COMMUNICATION

For historical reasons, we have two layers of communications in our system, both based on XML. The first layer, the oldest one, allows the communications between components of the middleware, or between the middleware and the various clients. It uses proprietary XML tags and no external references such as data types. This is mostly the consequence of early adoption of SGML since 1997, prior to the existence of any standards. The second layer is currently being implemented. It will be used first for communication with external systems and should progressively replace the first layer. It will be based on the HL7-CDA architecture and SOAP[2]. The HL7 Clinical Document Architecture is still in a transition period. CDA Release 1 became an ANSI-approved HL7 Standard in November 2000, and was the very first specification resulting from the HL7 Reference Information Model (RIM). Since then, the RIM has evolved, as well as the methodology used to derive RIM-based specifications. The main evolutionary steps in CDA Release 2 are that both header and body are now fully RIM-derived. In addition, there is now a richer collection of entries to use within CDA structures, such as enabling clinical content to be formally expressed to the extent that is it modeled in the RIM. We will, at term, use these specifications for all clinical document exchanges.

CONCLUSION

The patient record is made of very heterogeneous documents originating from numerous sources. Most of these documents are made both of structured data and narratives. Some of them, such as admission notes from general practitioners, will not be available largely in electronic form for a long time and must be scanned. We have developed a way to have a unique repository for all these documents and data in implementing a dual storage architecture, made of a patient-centric database that is file-oriented, and a data-centric relational database, which uses an attribute-value entity based representation that are tightly integrated. This dual storage architecture offers both the powerful representation capacity of EAV and the versatility of file based systems. It permits both a structured representation of content, whilst keeping the integrity of documents with their original layouts. It is part of our n-tiers component-based architecture and uses XML formatted messages that should evolve to the HL7-CDA standard in the near future. A manual analysis of 320 paper record allowed us to develop a preliminary simple ontology for documents types, while we are still working on a light-model shielding our clinical attribute dictionary. The separation of the semantic model and the attributes dictionary layer allows a fast growth of the number of data contexts, while the building of the semantic model continues at a slower speed. It remains an important problem and challenge that will have to be solved. HL7-CDA is a major step for exchanging clinical documents, but dedicated solutions must be found for storage and acquisition, especially when complex document layout is required. More than 7’000’000 documents are available in our computerized patient record using this architecture, around 5’000 are added every day.

REFERENCES

1. Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, et al. The HL7 Clinical Document Architecture. J Am Med Inform Assoc 2001;8(6):552-69.
2. Scherrer JR, Lovis C, Baud R, Borst F, Spahni S. Integrated computerized patient records: the DIOGENE 2 distributed architecture paradigm with special emphasis on its middleware design. Stud Health Technol Inform 1998;56:15-31.
3. Dolin RH, Alschuler L, S. B, Biron PV, Beebe C. HL7 Clinical Document Architecture. Release 2.0; 2003.
4. Tang PC, LaRosa MP, Gorden SM. Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. J Am Med Inform Assoc 1999;6(3):245-51.
5. Frazier P, Rossi-Mori A, Dolin RH, Alschuler L, Huff SM. The creation of an ontology of clinical document names. Medinfo 2001;10(Pt 1):94-8.
6. Lovis C, Baud RH, Revillard C, Pult L, Borst F, Geissbuhler A. Paragraph-oriented structure for narratives in medical documentation. Medinfo 2001;10(Pt 1):638-42.
7. Baud RH, Lovis C, Ruch P, Rassinoux A-M. A Light Knowledge Model for Linguistic Applications. Proc AMIA Annu Fall Symp 2001:37-41.
8. Chen RS, Nadkarni P, Marenco L, Levin F, Erdos J, Miller PL. Exploring Performance Issues for a Clinical Database Organized Using an Entity-Attribute-Value Representation. J Am Med Inform Assoc 2000;7(5):475-487.

Acknowledgements

Part of this work has been funded by the Swiss National Science Foundation **632 -066041**.

[2] Simple Object Access Protocol, www.w3.org/TR/SOAP/